

# 期末大作业说明

## 1 分组

1 - 2 位同学一组，自愿组队。组队信息请填写

● 【腾讯文档】统计学习方法2023期末大作业分组 <https://docs.qq.com/sheet/DWkpJR1J.NZFRmdW1l>

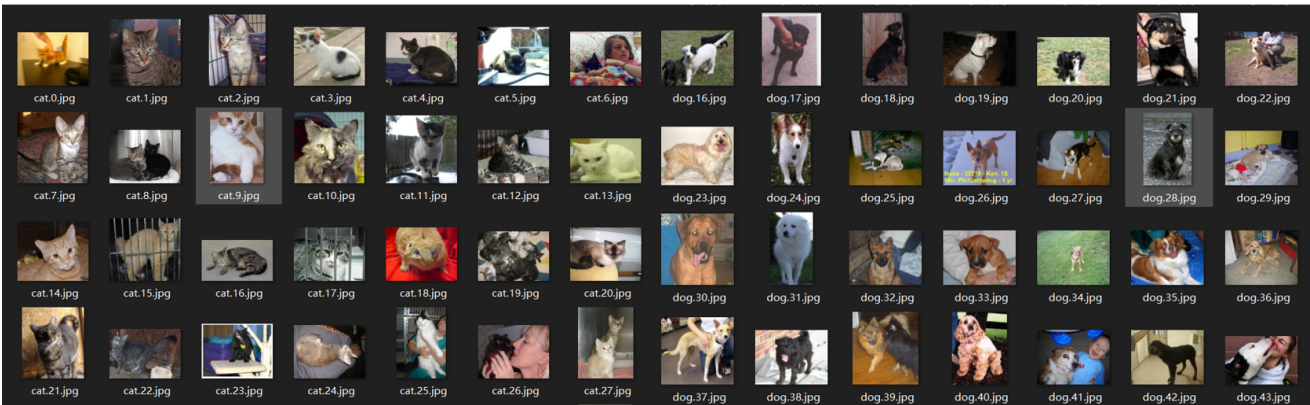
## 2 作业内容

有2个作业选题，请任选其一：①猫/狗分类任务 ②电子病历分类任务

### ① 猫/狗分类任务

编写一个算法程序来对图像中是否包含狗或猫进行分类。采用本次数据集，完成猫狗分类任务。

### 数据集



文件结构如下：

|                       |                 |         |        |
|-----------------------|-----------------|---------|--------|
| test                  | 2013/9/20 10:19 | 文件夹     |        |
| train                 | 2022/5/19 13:35 | 文件夹     |        |
| validation            | 2022/5/19 13:35 | 文件夹     |        |
| sample_submission.csv | 2019/12/11 4:18 | XLS 工作表 | 112 KB |

- 训练集（类别已在文件名标注）包含10000张dogs以及10000张cats
- 测试集（无标注）包含12500张dogs和cats图片
- 验证集（文件名已标注类别）包含2500张dogs和2500张cats图片。

数据集下载链接：

- 天翼网盘链接：<https://cloud.189.cn/t/6bmyMjAbY77f>（访问码：xo8m）

- 百度云盘链接: [https://pan.baidu.com/s/1qyjbasqKDBLr\\_m4L4nC9Gg](https://pan.baidu.com/s/1qyjbasqKDBLr_m4L4nC9Gg) (提取码: lrr6)

同学们注意解压缩的时候解压缩到文件夹哦! 不然就全解压缩到一个文件里去啦~




## 输入输出

要求根据最终的模型, 输出**测试集**中每张ID图片对应是**狗**的概率(0-1之间)。即0代表猫, 1代表狗。同学们可以针对模型, 计算验证集的准确率, 供同学们选择较好模型或调参。

## ② 电子病历分类任务

### 数据集

文件结构如下:

|   |                |                         |          |
|---|----------------|-------------------------|----------|
|  training_dataset    | 2023/6/4 15:30 | Microsoft Excel 逗号分隔值文件 | 6,225 KB |
|  test_dataset        | 2023/6/2 16:58 | Microsoft Excel 逗号分隔值文件 | 1,645 KB |
|  sample_submission_2 | 2023/6/4 15:39 | Microsoft Excel 逗号分隔值文件 | 240 KB   |

数据集内容如下:

|    | A     | B   | C   | D     | E   | F      | G   | H   | I    | J   | K     | L   | M     |
|----|-------|-----|-----|-------|-----|--------|-----|-----|------|-----|-------|-----|-------|
| 1  | 序号    | 过敏史 | 冠心病 | 冠心病年限 | 心绞痛 | 陈旧心肌梗死 | PCI | CAG | CABG | 高血压 | 高血压年限 | 糖尿病 | 糖尿病年限 |
| 2  | 16248 | 1   | 2   |       | 2   | 1      | 1   | 1   | 1    | 2   |       | 2   |       |
| 3  | 16249 | 2   | 1   | 2     | 1   | 1      | 2   | 2   | 1    | 1   | 30    | 2   |       |
| 4  | 16250 | 2   | 1   | 30    | 1   | 1      | 1   | 1   | 1    | 1   | 10    | 2   |       |
| 5  | 16253 | 0   | 0   |       | 0   | 0      | 0   | 0   | 0    | 0   |       | 0   |       |
| 6  | 16254 | 1   | 1   | 2     | 2   | 1      | 1   | 2   | 1    | 1   | 10    | 2   |       |
| 7  | 16256 | 1   | 2   |       | 2   | 1      | 1   | 1   | 1    | 1   | 2     | 2   |       |
| 8  | 16257 | 1   | 2   |       | 2   | 1      | 1   | 1   | 1    | 1   | 28    | 2   |       |
| 9  | 16258 | 2   | 1   | 8     | 1   | 2      | 2   | 2   | 1    | 1   | 6     | 2   |       |
| 10 | 16259 | 1   | 1   | 12    | 2   | 2      | 2   | 2   | 1    | 2   |       | 2   |       |

包含**23789**个电子病历样本, 其中**18789**个为包含标签的**训练集**样本, **5000**个为待预测标签的**测试集**样本。

每个样本中包含病人基础信息、病史、化验结果等137项病历信息。测试集仅包含病历信息, 训练集除了病历信息外, 最后一列“**label**”表示分类标签:

**0**: “正常”; **1**: “脑卒中”; **2**: “心肌梗死”; **3**: “心力衰竭”; **4**: “血运重建”; **5**: “心源性死亡”共6类。

## 输入输出

要求根据最终的模型, 输出**测试集**中每个病历样本分别属于**0-5**类的概率(0-1之间)。

## 3 汇报

第18周上课时, 以**小组**为单位准备PPT, 进行汇报, 每组时长约**5**分钟

# 4 提交

提交zip压缩包文件，至 <http://xzc.cn/G3lYln4a84>

- submission.csv 文件（单独，作为评分依据）
- 代码文件
- 作业报告（详见模板文件，且需要写小组成员的分工情况）

● submission.csv 文件：测试集预测结果。

## 任务1 猫狗分类

样例见sample\_submission\_1.csv。

id为测试集图片id，label表示每张ID图片对应是狗的概率(0-1之间)。

|    | A  | B     | C   |
|----|----|-------|-----|
| 1  | id | label |     |
| 2  |    | 1     | 0.5 |
| 3  |    | 2     | 0.5 |
| 4  |    | 3     | 0.5 |
| 5  |    | 4     | 0.5 |
| 6  |    | 5     | 0.5 |
| 7  |    | 6     | 0.5 |
| 8  |    | 7     | 0.5 |
| 9  |    | 8     | 0.5 |
| 10 |    | 9     | 0.5 |
| 11 |    | 10    | 0.5 |

## 任务2 电子病历分类

样例见sample\_submission\_2.csv

序号为测试集病人序号，B-G列表示每个病历样本分别属于0-5类的概率(0-1之间)

|    | A     | B      | C      | D      | E      | F      | G      |
|----|-------|--------|--------|--------|--------|--------|--------|
| 1  | 序号    | 0      | 1      | 2      | 3      | 4      | 5      |
| 2  | 16246 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 3  | 16252 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 4  | 16268 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 5  | 16274 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 6  | 16284 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 7  | 16302 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 8  | 16304 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 9  | 16315 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 10 | 16316 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 11 | 16325 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 12 | 16332 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 13 | 16353 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 14 | 16357 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 15 | 16358 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 16 | 16377 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 17 | 16403 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 18 | 16407 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 19 | 16459 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |
| 20 | 16480 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 | 0.1667 |

## 文件命名方式

填写小组编号和小组成员姓名(用“\_”间隔)

若选题为“猫/狗分类任务”命名为“1.zip”；若选题为“电子病历分类任务”命名为“2.zip”

## 截止时间

2023-07-02(第18周 周日晚)

期末大作业

提交zip压缩包:

• submission.csv文件 (单独, 作为评分依据)

• 代码文件

• 作业报告

文件命名方式:  
填写小组编号和小组成员姓名(用“\_”间隔)  
若选题为“猫/狗分类任务”命名为“1.zip”; 若选题为“电子病历分类任务”命名为“2.zip”

截止时间  
2023-07-02(第18周 周日晚)

文件提交区域

小组编号:

小组成员姓名:

拖拽文件上传  
(或点击)

最多上传20个文件, 单个文件最大20M

提交

由快速创建收件夹提供技术支持

## 5 最终评分

模型得分(根据以下**loss**计算)+算法评测+贡献度+报告

根据提交文件中预测的概率，通过以下损失函数来判断模型得分：

$$\mathcal{R}(\mathbf{W}) = -\frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C y_c^{(n)} \log \hat{y}_c^{(n)}$$

其中，N为测试集图片数目，c表示类别数目，y\_hat为提交文件中的预测概率，y为ground\_truth(真实标签)。